**Objectives**

**Methods**

**Metagenomic Binning Procedure**

The requisite numerical information recommended by the Genomic Standards Consortium for metagenome-assembled genome (1) is shown in Table X. The following approaches apply to all bins derived. Taxonomic classification was performed using the taxator-tk algorithm (). Reassembly of bins and initial co-assembly was performed using SPAdes 3.11(). The initial assembly was completed in metagenomic mode and reassembly was done for each bin in `careful` mode using the first pass contigs as `untrusted contigs` .

Contigs were binned according to their coverage & tetramer frequencies. A set of consensus bins were derived from the bins produced by the maxbin2, metabat2, and concoct algorithms. Completeness and contamination assessment were performed using the lineage workflow in CheckM. The bioinformatics pipeline from QC, to assembly, to binning, to refinement, to reassembly, and taxonomic classification was done within the metaWRAP software. Prokka () was used to facilitate gene calling and preliminary annotations. Within prokka, Prodigal, barrnap, and Aragorn v1.2 were used to call ORFs, rRNA & tRNA respectively (,,). All protein sequences were generated using Translation Table 11.

**Metabolic Model Processes**

Additional protein annotations were conducted using `metabolic-hmms` collection provided by the Banfield lab (), the ` dbCAN-fam-HMMs.v6` collection provided by BioEnergy Science Center of the DOE (), and the KEGG BlastKOALA and GhostKOALA annotation web service ().

**Linking OTUs from Shotgun Libraries to Bins & Amplicon Profiles**

Both OTUs derived from shotgun libraries, as well as those observed in amplicon libraries prepared for samples collected on 2013-08-12 were assigned to bins based on similarities in abundance and taxonomic classification. Replicate samples were combined by their mean. The minimum observed Pearson correlation between replicates was 0.79. Amplicon OTUs were filtered by the sum of their abundances in comparable samples to the top 500/1000. A total of 11 samples were compared between library types including odd depths from 3 to 17 and 20, 21, and 22 meters. Abundances were normalized using the L1 norm within each sample and then within each OTU.

Raw abundance (*A*) of a bin in a given sample *j* were calculated within metaWRAP using the following equation:

The product of the length of contig (*li*) and its coverage (*cij*) in sample *j* was summed across all *N* contigs assigned to the bin and then divided by total length of the bin genome. The raw abundances in each sample was divided by the number of reads in each library before following the same normalization steps applied to OTU abundances.

OTU taxonomy was assigned using the RDP classifier using the 16s rRNA Training Set 16. Confidence scores below 50% and *Incertae Sedis* classifications were removed from the “fixed rank” output. Each level in the hierarchy between Kingdom & Genus was treated as an independent feature to match the format of the taxonomic assignments produced by metaWRAP.

Pairwise Euclidean distances between abundance vectors for bins and OTUs ranged from 0 to 1. The fraction of matching taxonomic labels was then subtracted from the distance to produce a combined distance metric ranging from 1 to -1, where the latter represents a perfect match. The tolerance of this metric for producing correct matches was a challenge to assess with a relative dearth of known matches usable for scoring assignments. 8 bins were used for this purpose as these 5 of them received an assembled 16S rRNA gene copy and the other three were positive control organisms. The tolerance was trained on these bins and was observed to be different for the amplicon library data (0.0251), as compared to shotgun library data (0.176). The preprocess and assignment routines are outlined in the `assign\_bin\_16s.py` script and in annotated form in the `16S\_to\_Bin\_Matching` jupyter notebook.

**Gene Abundance Quantification**

The abundances for select genes was annotated with biogeochemical process-related KOs. The strict nucleotide sequence of each gene was bookended by 50-100 bp on each end to ensure that near identical gene sequences were diluted using contig-specific surrounding sequence. Salmon was used to quantify the abundances of these sequences in each library.

1. Robert M Bowers, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T B K Reddy, Frederik Schulz, Jessica Jarett, Adam R Rivers, Emiley A Eloe-Fadrosh, Susannah G Tringe, Natalia N Ivanova, Alex Copeland, Alicia Clum, Eric D Becraft, Rex R Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M Weinstock, George M Garrity, Jeremy A Dodsworth, Shibu Yooseph, Granger Sutton, Frank O Glöckner, Jack A Gilbert, William C Nelson, Steven J Hallam, Sean P Jungbluth, Thijs J G Ettema, Scott Tighe, Konstantinos T Konstantinidis, Wen-Tso Liu, Brett J Baker, Thomas Rattei, Jonathan A Eisen, Brian Hedlund, Katherine D McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W Tyson, Christian Rinke, The Genome Standards Consortium, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H Parks, A Murat Eren, Lynn Schriml, Jillian F Banfield, Philip Hugenholtz & Tanja Woyke. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology 35(725–731) doi:10.1038/nbt.3893
2. Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726-731 (http://www.kegg.jp/ghostkoala/)